معمارية الحاسوب

Architecture Computer



Memory & Cache Memory

الذاكرة وذاكرة السريعة

Characteristics
الخصائص

- الموقع (Location)
- السعة (Capacity)
- وحدة النقل (Unit of transfer)
- طريقة الوصول (Access method)
- الاداء (Performance)
- النوع المادي (Physical type)
- الخصائص الفيزيائية (Physical characteristics)
- التنظيم (Organisation)

Dr. ramzi elghanuni_Lecture9

Location

الموقع

- CPU
- Internal
- External

Cache Memory

عبارة عن ذاكرة صغيرة الحجم ,سرعتها عالية جدا ,تكون بين CPU وبين RAM الهدف منها هي مجاراة سرعة المعالج التي تكون سرعتها عالية جدا.

لا يمكن مجاراة سرعة CPU من خلال الذاكرة الرئيسية (RAM) نضع (Cache Memory) بينهم بحيث تكون سرعتها متوسطة بين سرعة CPU وسرعة RAM.

Capacity
السعة

السعة التخزينية Cache Memory نحددها من خلال Word size

كل Word size تحتوي على مجموعة من Bits.

$$N = 2^A$$

N عدد Word size الموجودة في Cache

$$A = 8 \ bit$$

A عبارة عن عدد Bits لكل موقع تخزيني (عدد المواقع التخزينية)

$$N = 2^8 = 256 \ \text{Address}$$

أول موقع 00000000

ثاني موقع 00000001

في حالة : $A = 16 \ bit$ $N = 2^{16} = 65535$

Dr. ramzi elgharmuni_Lecture5

Unit of Transfer
وحدة النقل

- وحدة نقل البيانات في الذاكرة الرئيسة (Main Memory) هي عدد bits التي يتم قراءتها أو كتابتها في الذاكرة في وحدة الزمن.
- وحدة النقل (Unit of Transfer) مرتبطة بعرض الناقل البيانات Data Bus.
- وحدة العنونة (Addressable unit) بناء على أصغر موقع تخزيني.

Access Methods (1)
طرق الوصول

- **المتسلسل (Sequential)**

التسلسل يكون من البداية حتي الوصول للبيانات ، يتم التسلسل في عملية البحث عن البيانات حتي نصل إلي البيانات المطلوبة ، Access Time زمن الوصول يعتمد على مكان وجود البيانات مثال عليها الشريط المغناطيسي (tape) وتكون سعة التخزين عالية جدا.

- **المباشر (Direct)**

تكون الذاكرة مقسمة إلي blocks يتم الوصول إلي أي blocks منها من خلال الوصول إلي منطقة مجاورة من خلال القفز ولها عنوان خاص بها سرعة الوصول تعتمد على الموقع التخزيني للبيانات والموقع السابق ,مثال عليها (Hard disk) سرعة الوصول وسعة التخزين كبيرة.

Dr. ramzi elghanuni_Lecture9

Access Methods (2)
طرق الوصول

- العشوائي (**Random**)

يكون الوصول عشوائي للمواقع وزمن الوصول لا يعتمد على الموقع التخزيني ومثال عليها RAM.

- الترابطي (**Associative**)

تكون البيانات موجودة من خلال مقارنة جزء من المحتوي مع المخزن ، نقارن البيانات الجديدة بالقديمة فالبيانات تكون مترابطة مع بعضها البعض وزمن الوصول يكون مستقل عن المكان التخزيني أو الوصول السابق ومثال عليها cache إذا تكون البيانات مترابطة مع بعضها البعض.

Dr. ramzi elghanuni_Lectures

Memory Hierarchy

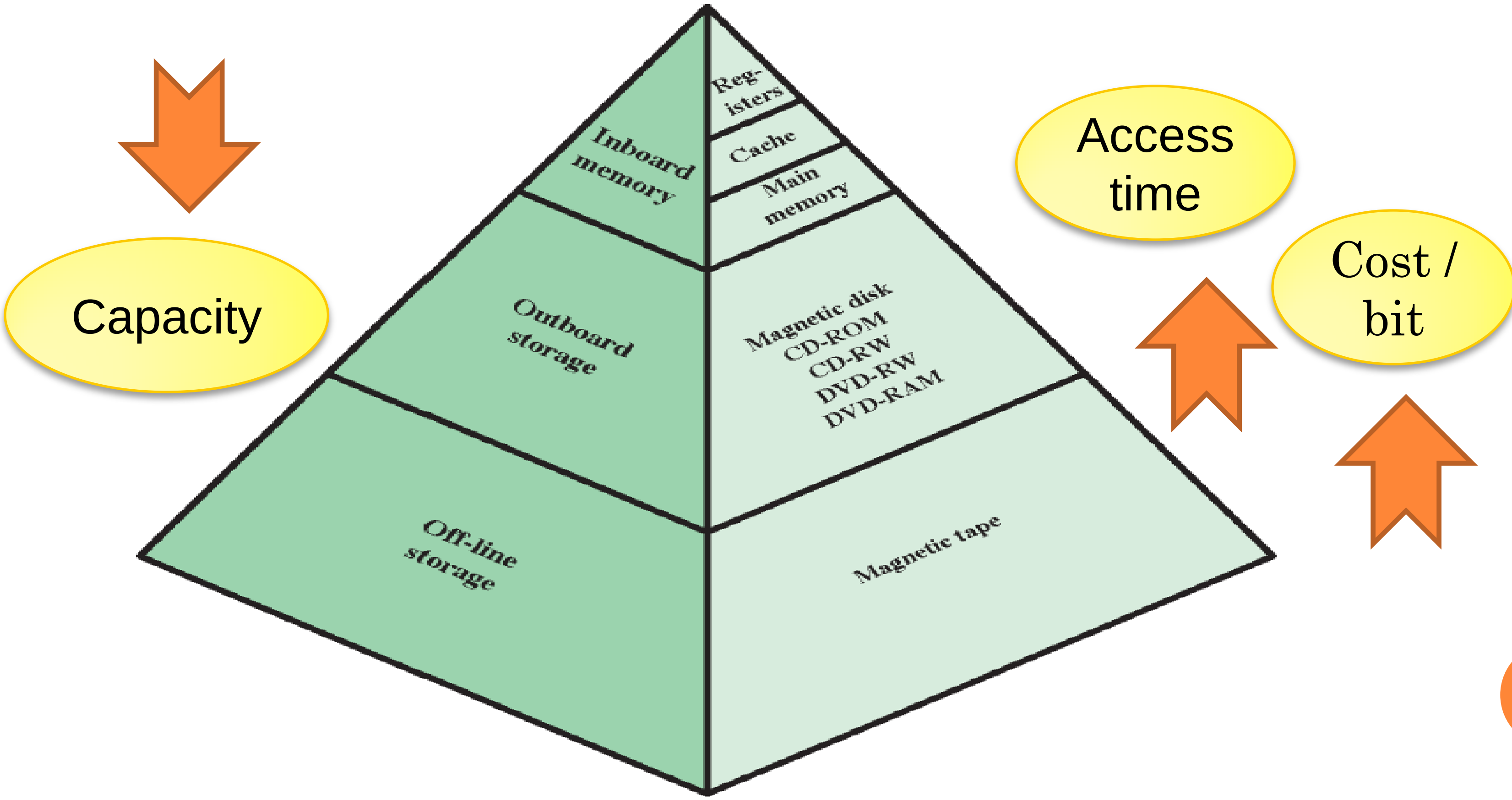
التسلسل الهرمي للذاكرة

تدرج Memory بناء على الاكثر سرعة :-

- **Registers**
  - In CPU
- **Internal or Main memory**
  - May include one or more levels of cache
  - “RAM”
- **External memory**
  - Hard Disk

Memory Hierarchy – Diagram

مخطط ـ التسلسل الهرمي للذاكرة



Dr. ramzi elghanuni_Lecture9

Performance
الاداء

- زمن الوصول (Access time)

زمن تقديم البيانات والحصول عليها في Registers أعلي من Cache و Cache أعلى من RAM.

- زمن دورة الذاكرة (Memory Cycle time)

زمن الوصول للبيانات بالإضافة إلي استردادها (أي زمن مطلوب قبل بداية الوصول التالي)

Cycle time access + recovery

- معدل النقل (Transfer Rate)

نسبة نقل البيانات من مكان إلي مكان , وسرعة هذا النقل.

Dr.ramzi alghanuni_Lecture9

Physical Types
النوع المادي

- Semiconductor
  - RAM
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD
- Others
  - Bubble
  - Hologram

Physical Characteristics
الخصائص الفيزيائية

- ✓ الاضمحلال (Decay) تضاؤل الاداء أو السعة .
  - في الذاكرة المتطايرة المعلومات تضمحل بشكل طبيعي أو يتم فقدانها عند غلق / أنهاء الطاقة الكهربائية.
  - في الذاكرة الغير متطايرة المعلومات المسجلة لا تفقد أو تضمحل حتي يتم تغييرها بشكل معتمد ولا تحتاج إلي الطاقة الكهربائية للحفاظ على المعلومات.
- ✓ استهلاك الطاقة (Power consumption).
- ✓ تطاير البيانات (Volatility).
- ✓ قابلة للمسح (Erasable) هل البيانات قابلة للمسح أو لا.

الذاكرة المغناطيسية غير متطايرة أما ذاكرة أشباه الموصلات (الالكترونية) قد تكون متطايرة أو غير متطايرة.

Dr. ramzi elghanuni_Lectures

Organisation

التنظيم

ترتيب الذواكر يكون للمواقع التخزينية حسب عدد Bits ودائما لا يكون واضح فمثلا Hard Disk يكون في فواصل بينية.

عملية المقارنة بين الذواكر

- How much?
  - Capacity
- How fast?
  - Time is money
- How expensive?

Hierarchy List

قائمة التسلسل الهرمي

الترتيب من حيت السرعة :-

- Registers
- L1 Cache
- L2 Cache
- Main memory
- Disk cache
- Disk
- Optical
- Tape

So you want fast?

سؤال:

هل نستطيع حذف (RAM) ونضع مكانها (Cache Memory) ولأدعي أن نحتاج إلي RAM ، يعني مباشرة من Cache Memory التي تكون Static RAM مباشرة إلي CPU تكون السرعة عالية جدا وكذلك التكلفة تكون عالية جدا لذلك نطرا إلي إستخدام RAM لتقليل تكلفة تصنيع الكمبيوتر.

Dr. ramzi elghanuni_Lecture9

Locality of Reference
مرجعية المحل

تعمل Cache Memory بمبدأ مرجعية المحل.

البيانات التي تمت الاشارة إليها مسبقا في الذاكرة الرئيسية تميل إلي إمكانية الاشارة إليها مرة اخري في وقت لاحق. بمعني أن البيانات التي نستخدمها في Cache Memory يمكن أن استخدامها مرة اخري.

يوجد استراتيجيتين :-

- Temporal Locality
- Special Locality

Dr/ramzi elghanuni Lectures

Locality of Reference

مرجعية المحل

➤ **Temporal Locality**

اخر تعليمة تم استدعاها من RAM من المحتمل أن يتم استدعائها مرة اخرى فيتم الاحتفاظ بنسخة من هذه التعليمة لوقت معين إذا لم يتم استخدامها يتم التخلص منها.

➤ **Special Locality**

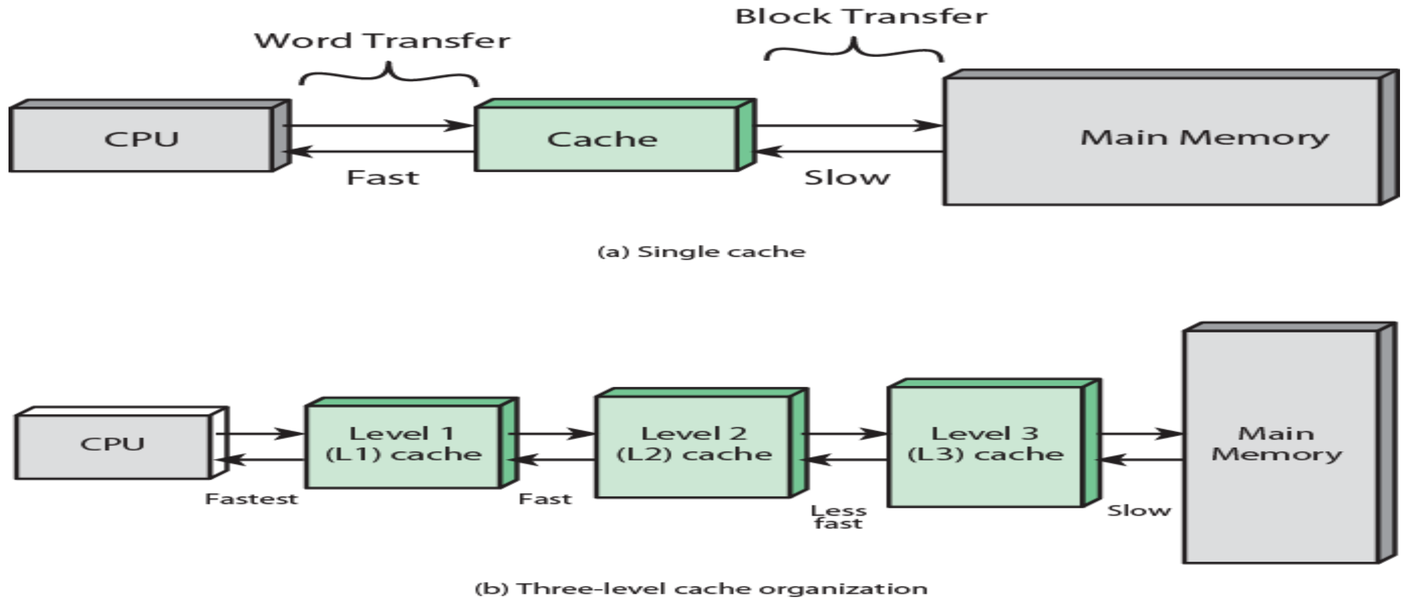
التعليمة الموجودة من القرب من التعليمة التي نريد استدعاه من RAM يمكن أن نستخدمها بدل من جلب تعليمة واحدة فقط فيتم جلب blocks كامل أو جزاء كامل من RAM بحيث يمكن أن استخدم التعليمات التي بجنبه.

Dr. ramzi_elghanuni_Lecture9

Cache and Main Memory

Cache Memory عبارة عن ذاكرة صغيرة الحجم وسريعة جدا توضع بين RAM وبين CPU ويمكن أن توضع على CPU.

يتم استخدام النوع الثاني من مرجعية المحل (**Special Locality**) الذي ينقل blocks كامل من Main Memory إلي Cache وينقل Word من Cache إلى CPU على حسب الاستخدام و الطلب.



(a) Single cache

(b) Three-level cache organization

Dr. ramzi elghanmi_Lecture9

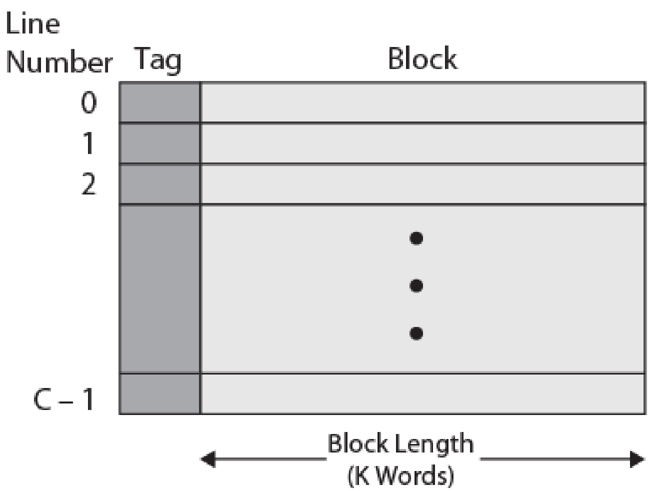
Cache operation – overview
العمليات على الذاكرة السريعة – نظرة عامة

- CPU يطلب محتويات موقع معين موجود في الذاكرة.
- Cache يقوم بفحص إذا كانت هذه البيانات موجوده لديه أو لا.
- إذا كانت البيانات موجودة لديه يرسلها مباشرة إلي CPU بهذا يعطينا سرعة عالية.
- إذا كانت هذه البيانات غير موجودة في Cache يقوم باستقبال blocks من الذاكرة يحتوي على المعلومة المطلوبة ثم يقوم بإرسال هذه المعلومة إلي CPU.

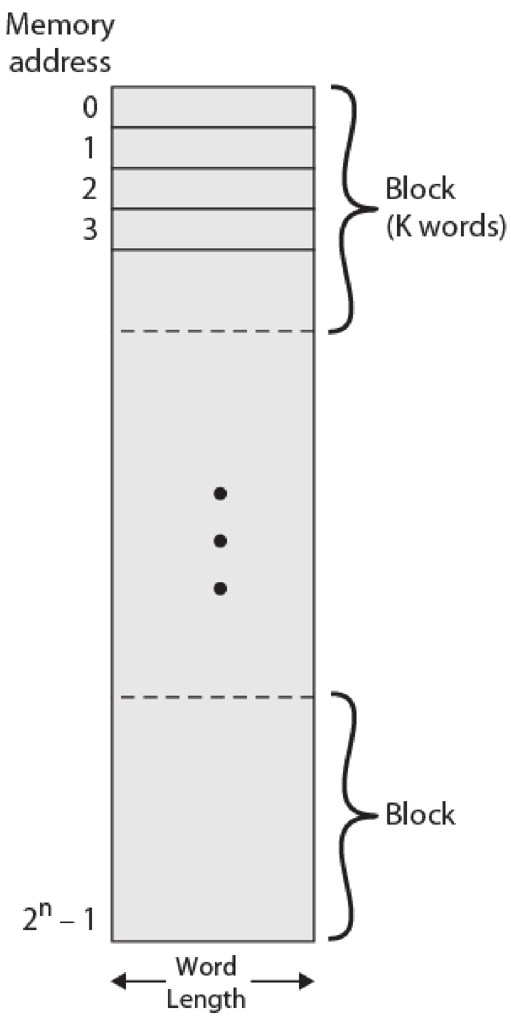
Dr. ramzi elghanuni_Lectures9

Cache/Main Memory Structure

بنية نظام الذاكرة الرئيسية والذاكرة السريعة



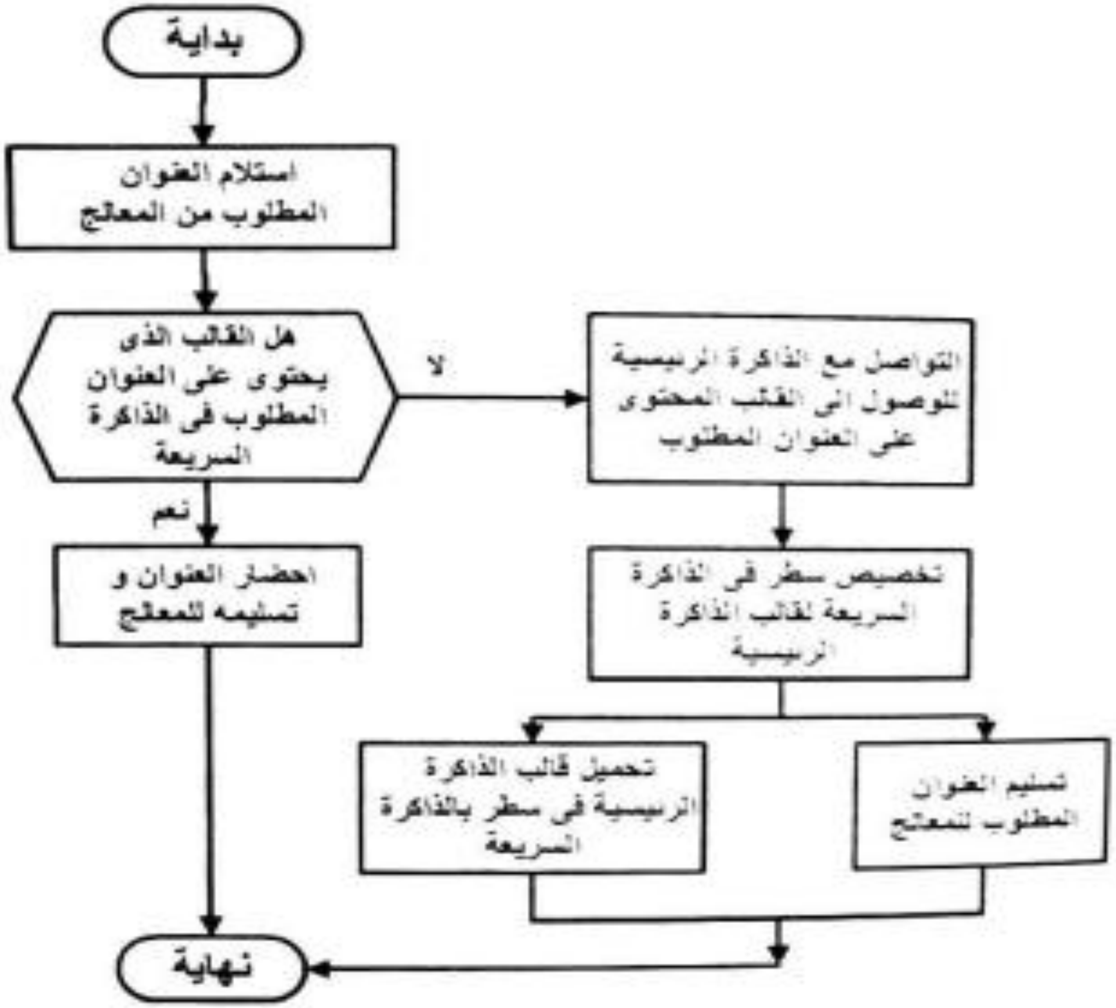
(a) Cache



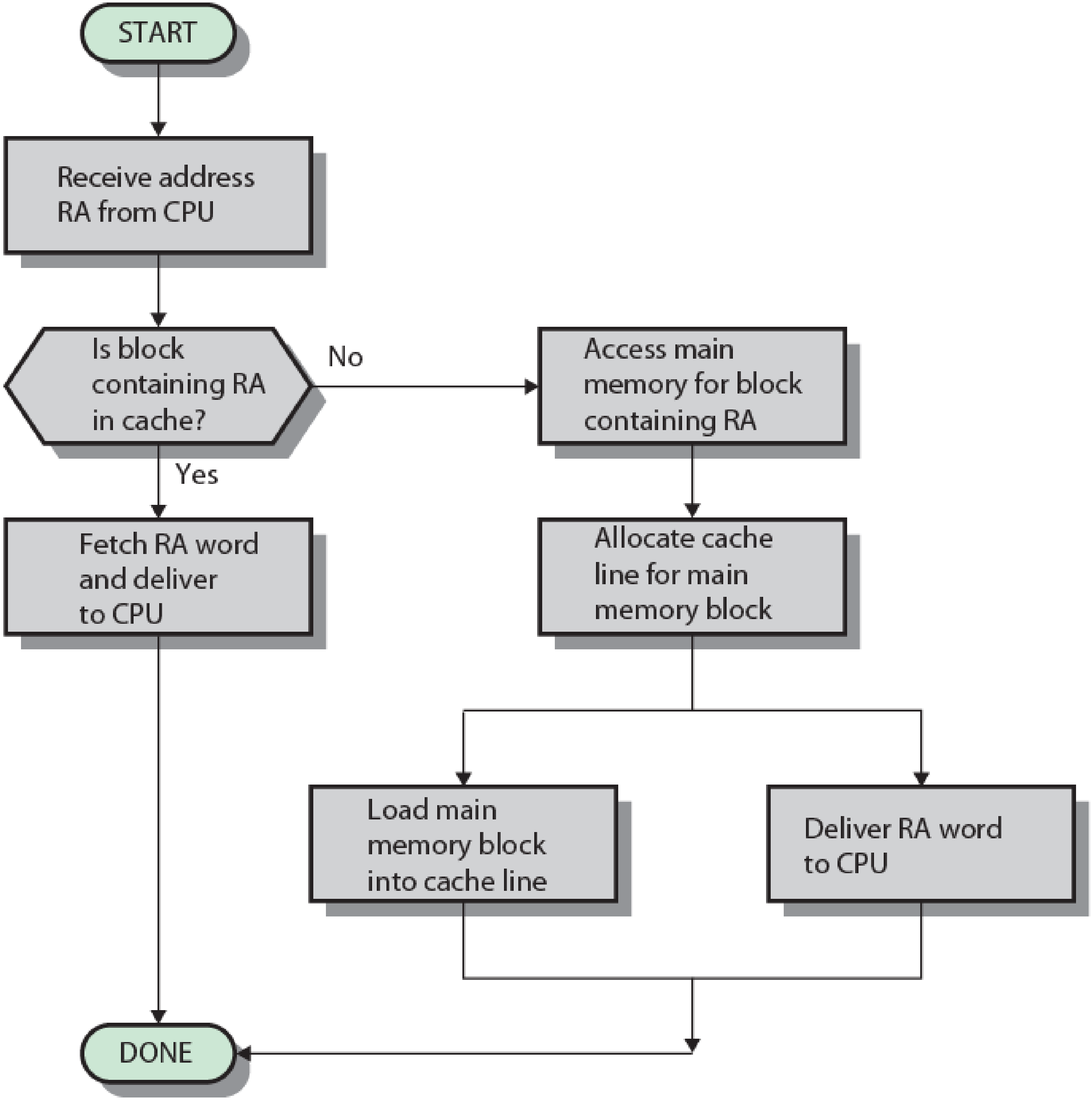
(b) Main memory

Cache operation – overview
العمليات على الذاكرة السريعة – نظرة عامة

- Cache يحتوي على tags.
- هذا tags يحتوي على معلومة تدل على اسم blocks الذي تم جلبه من RAM.
- كل معلومة موجودة في Cache تحتوي على tags



الشكل (7.5) – عملية القراءة من الذاكرة السريعة


START
Receive address RA from CPU
Is block containing RA in cache?
No
Yes
Access main memory for block containing RA
Allocate cache line for main memory block
Fetch RA word and deliver to CPU
Load main memory block into cache line
Deliver RA word to CPU
DONE

Cache Design
تصميم الذاكرة السريعة

- Addressing (logial,physical)
- Size
- Mapping Function
- Replacement Algorithm
- Write Policy
- Block Size
- Number of Caches

كل هذه العناصر تتحكم في أداء الكمبيوتر وسرعته وتكلفته.

Size does matter

مسالة الحجم

Cost

كلما زاد عدد Cache أو حجمها بتزيد التكلفة (Cost).

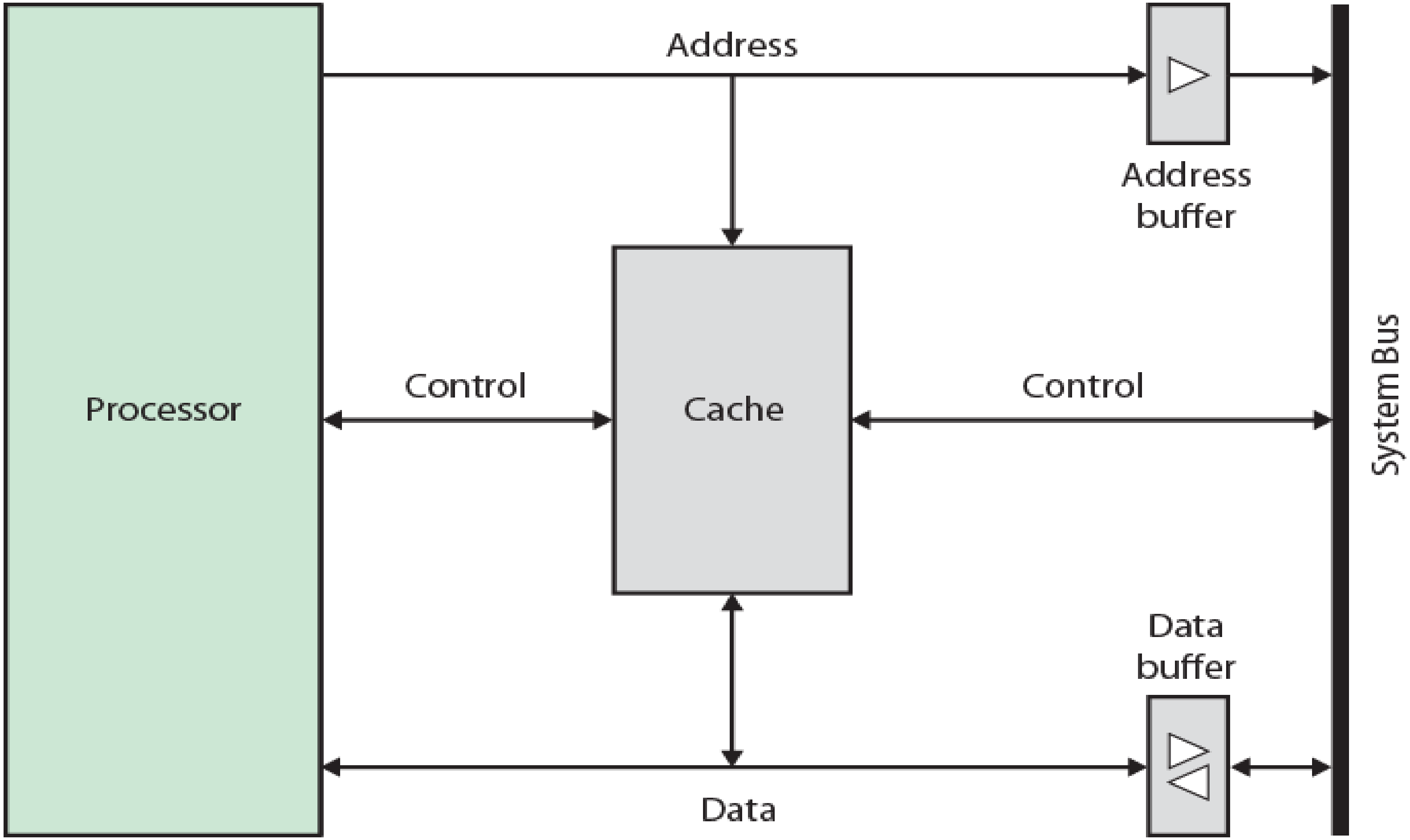
Speed

كلما زاد عدد Cache أو حجمها بتزيد السرعة (Speed) العملية طرديا.

Dr. ramzi elghanuni_Lectures

Typical Cache Organization

التنظيم النموذجي للذاكرة السريعة



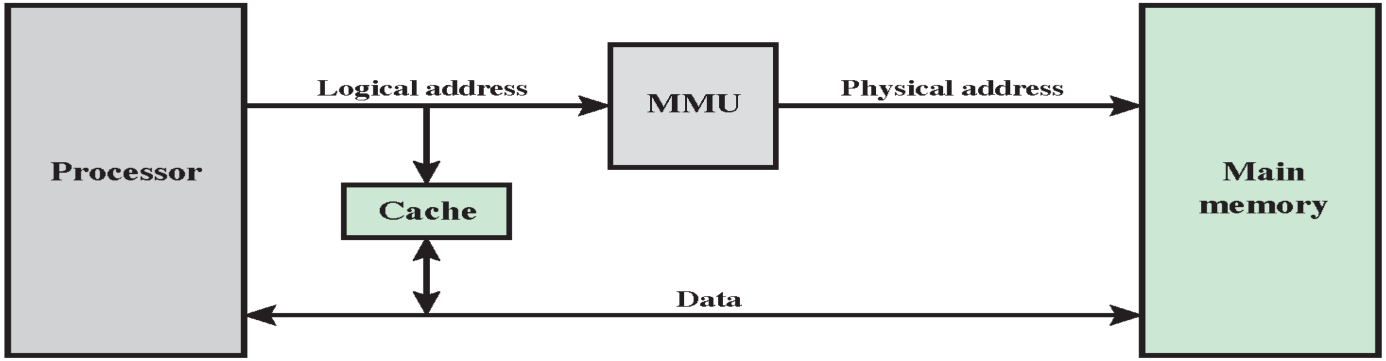
Addressing cache memory (1)
عناوين الذاكرة السريعة

عند استخدام عناوين افتراضية يختار المصمم نظام يضع الذاكرة السريعة بين المعالج و وحدة إدارة الذاكرة (MMU) أو بين وحدة إدارة الذاكرة و الذاكرة الرئيسية .

الذاكرة السريعة الظاهرية تخزن البيانات باستخدام العناوين الافتراضية.

المعالج يصل إلى الذاكرة السريعة مباشرة دون المرور عبر وحدة إدارة الذاكرة .

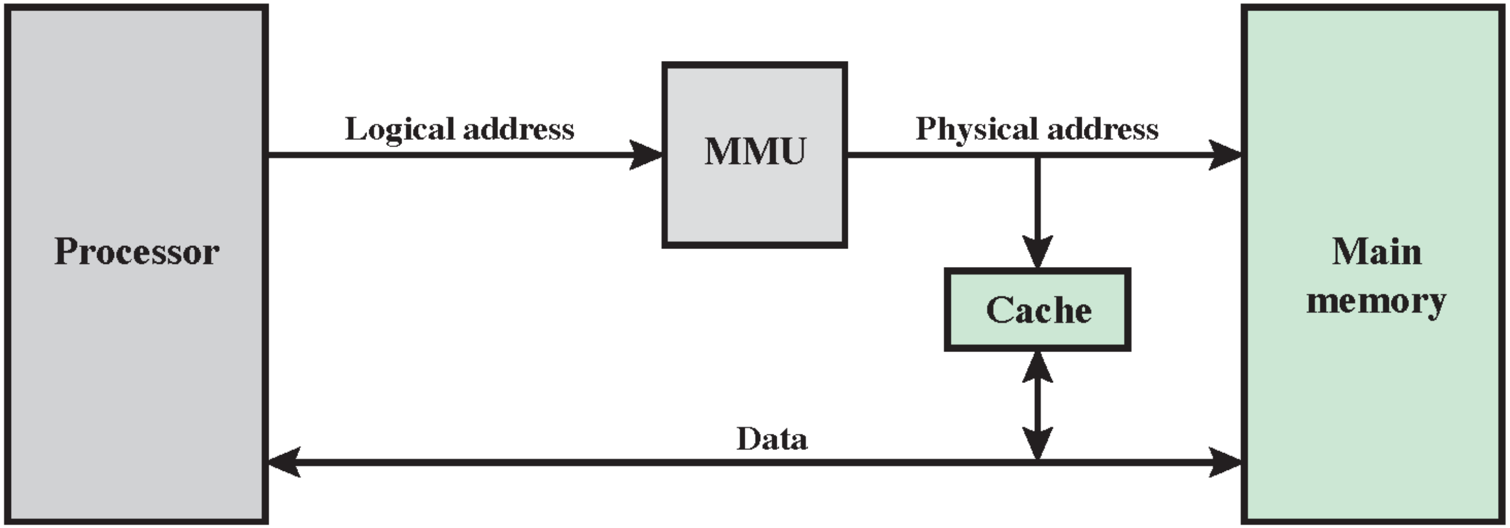
Dr. ramzi elghanuni_Lecture9



(a) Logical Cache

Addressing cache memory (2)
عناوين الذاكرة السريعة

الذاكرة السريعة المادية تخزن البيانات باستخدام عناوين الذاكرة الرئيسية الحقيقية.



(b) Physical Cache

Dr. ramzi elghanuni_Lecture9

size cache memory

حجم الذاكرة السريعة

يجب أن يكون حجم الذاكرة السريعة صغير بما فيه الكفاية بحيث يكون متوسط التكلفة الإجمالية للخانة (البت) هو قريب منه في الذاكرة الرئيسية فقط ، وكبيرة بما فيه الكفاية بحيث يكون إجمالي متوسط زمن الوصول قريبا من الذاكرة السريعة لوحدها .

هناك دوافع أخرى عديدة للتقليل من حجم الذاكرة السريعة منها أن كبر الحجم قد يقلل من السرعة و كذلك المساحة المتاحة على رقاقة المعالج محدودة.

Dr. ramzi elghanuni_Lecture9

Mapping Function (1)
وظيفة الاسقاط

- Blocks in cache called line
- كل مجموعة مواقع تخزينية في Main Memory تسمي Block
- نفس المجموعة في Cache تسمي Line
- بحيث يكون حجم Block يساوي حجم Line
- لو نريد نقل محتويات بيانات Blocks من Main Memory نريد نقلها علي Line يكون نفس الحجم.
- Line هو Blocks لكن التسمية تختلف في Cache Memory

Dr. ramzi elghanuni_Lecture9

Mapping Function (2)

وظيفة الاسقاط

- **Hit** البيانات موجودة في Cache بمعني اخر إذا قمنا بعملية بحث عن البيانات في Cache وقمنا بإيجادها.
- **Miss** إذا كانت البيانات ليست موجودة في Cache نظرا لجلبها من Memory Main

Direct Mapping

الاسقاط المباشر

الاسقاط المباشر طريقة لجلب البيانات من ذاكرة RAM إلي Cache

يتم تحميل كل (Block) من الذاكرة الرئيسية في سطر محدد من الذاكرة السريعة.

Dr. ramzi elghanuni_Lecture9

مزايا الاسقاط المباشر

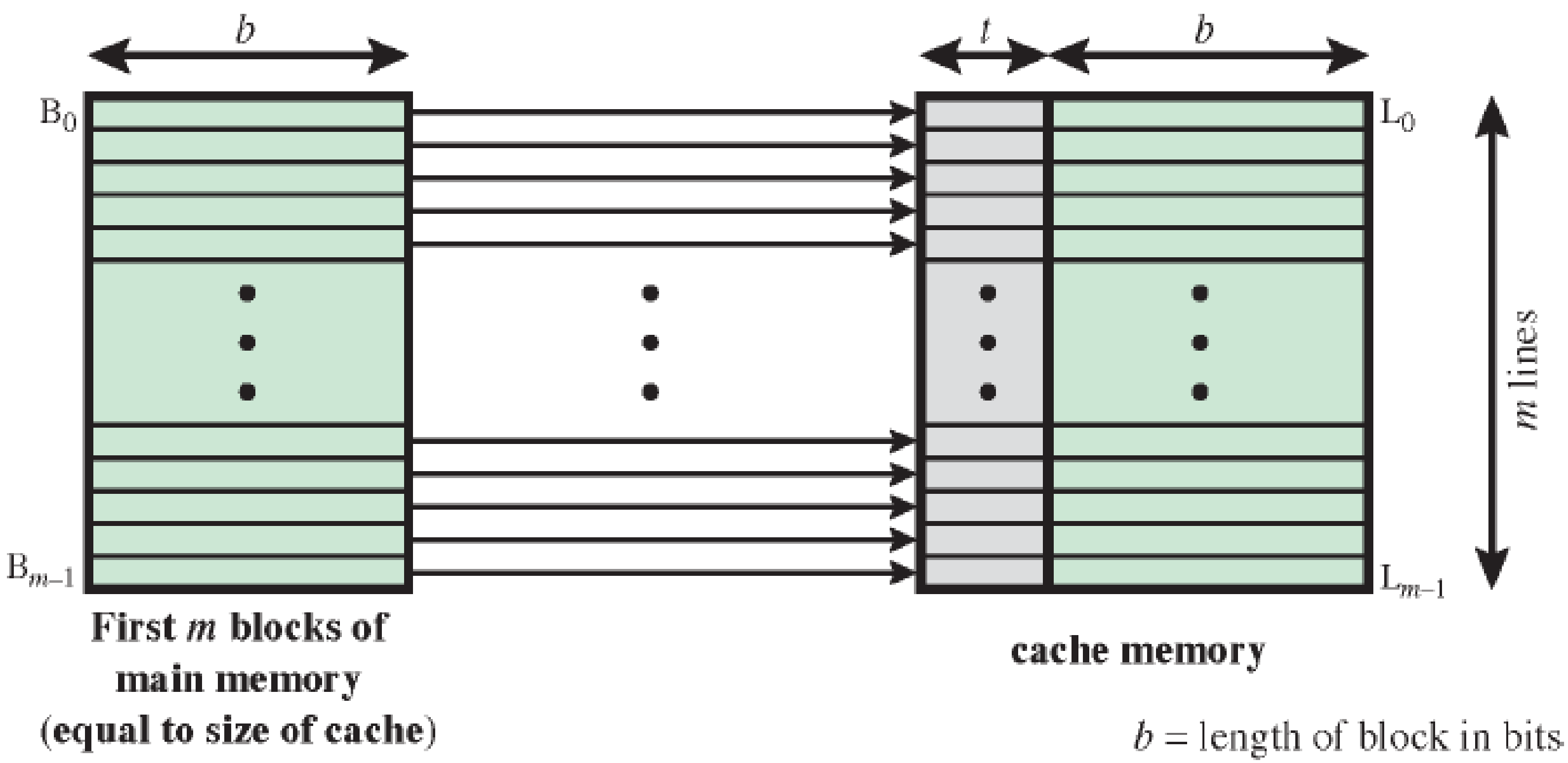
- أبسط تقنية معروفة.
- غير مكلفة بالنسبة لقدرات النظام.

عيوب الاسقاط المباشر

- للبيانات مكان واحد فقط في الذاكرة السريعة في حالة جلب بيانات جديدة وتكون متتابعة يلزم جلب كل مرة بيانات واحدة و إزالتها ثم وضع الأخرى وهكذا.

The mapping of the main memory to the cache memory: Direct

الاسقاط من الذاكرة الرئيسية الى الذاكرة السريعة : المباشر



(a) Direct mapping

Dr. ramzi elghanuni_Lecture9

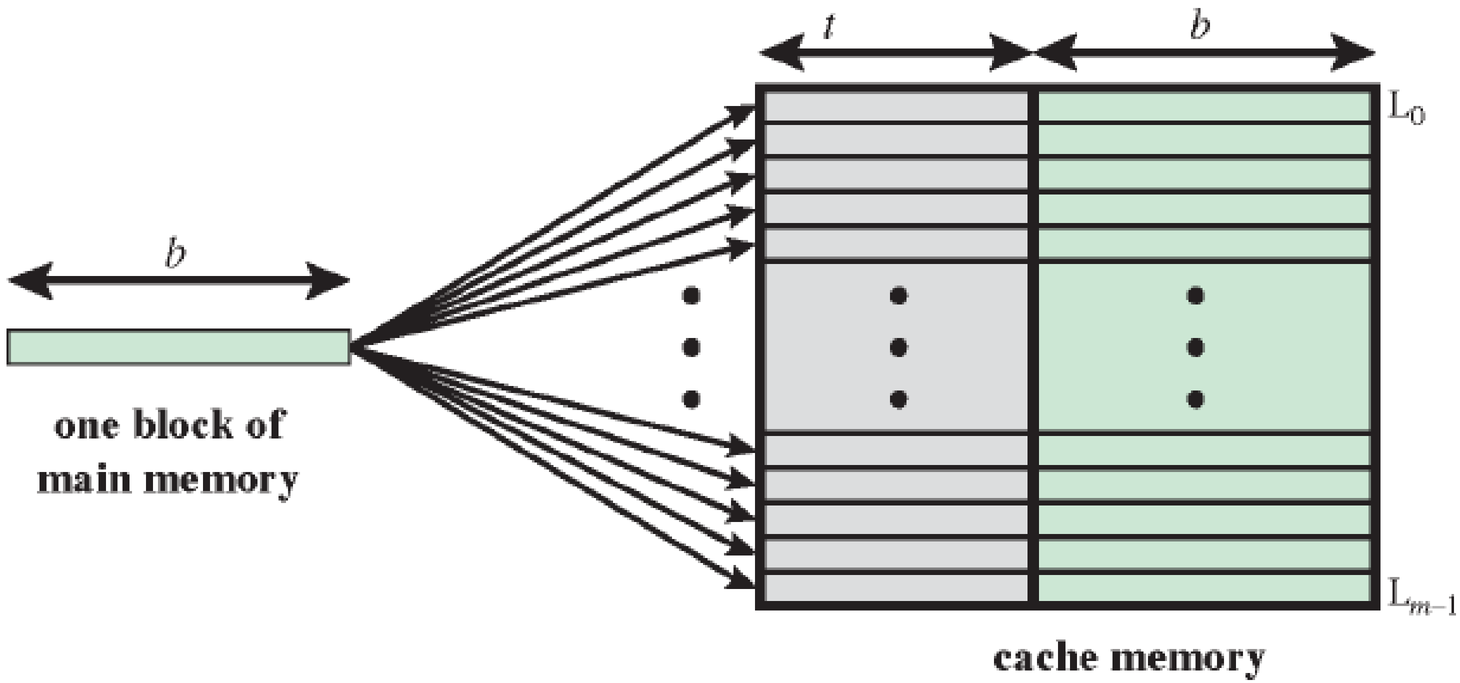
Associative Mapping

الاسقاط الترابطي

- يمكن وضع blocks في إي Line.
- المشكلة في عملية البحث يجب البحث في جميع Line.

The mapping of the main memory to the cache memory: Associative

الاسقاط من الذاكرة الرئيسية الى الذاكرة السريعة : الترابطي



(b) Associative mapping

Replacement Algorithms (1)
Direct mapping
خوارزميات الاستبدال
الاسقاط المباشر

Dr. ramzi elghanuni_lecture9

- عملية التبديل بين blocks الموجودة في الذاكرة.
- نستخدم خوارزميات التبديل (Replacement Algorithms).
- في الاسقاط المباشر (Direct mapping) لا يلزم أن نستخدم هذه الخوارزميات لأنه يتم استبدال blocks الجديد بـ blocks القديم.

Replacement Algorithms (2)
Associative mapping
خوارزميات الاستبدال
الاسقاط الترابطي

نستخدم هذه الخوارزميات لمعرفة أي Line سيتم وضع blocks الجديد فيه وإزالة القديمة.

كل Line موجودة في cache مرشحة للاستبدال توجد عدة طرق :-

- **Least Recently used (LRU)**

أقل من أستخدم أخير ،أطول فترة ممكنة ولم يستخدم بعد، يتم استبداله.

- **First In First Out (FIFO)**

أول من يدخل هو أول من يخرج.

- **Least frequently used**

أقل من تكرر استخدامه، يعني أقل blocks تم اختياره في التنفيذ.

- **Random**

وهي الاسواء.

Write Policy
سياسة الكتابة

<u>استراتيجيات أو سياسات الكتابة على Cache</u>

يجب التخزين في Memory وcache في نفس الوقت عند إدخال البيانات من I/O فيتم الكتابة في Memory و يتم الكتابة في cache في نفس الوقت لكي لا يحدث تعقيد ومن المهم أن يتم تحديث البيانات باستمرار في الذاكرة، يجب التأكد قبل الاستبدال أنه لا يوجد تعديلات في الذاكرة بحيث يجب أن يكون عنوان blocks وLine يحملن نفس القيمة أو العنوان.

Dr. Ramzi elghanuni_Lecture9

Write through
الكتابة من خلال

- الكتابة من خلال (**Write through**)

يتم كتابة البيانات من I/O إلـي cache وRAM وأي تحـديث علـى البيانـات فـي cache يتم بإرسال التحديث إلي RAM لذلك السرعة بطيئة وTraffic عالي.

Write back
إعادة الكتابة

- إعادة الكتابة **(Write back)**

يتم كتابـة البيانـات من I/O إلـي cache وRAM بعد انتهـاء التحـديث بشـكل كامـل فـي cache يتم إرسال التحديثات مرة واحدة فقط إلى RAM ، لا يسبب أي Traffic عالي.

Dr_ramzi elghanuni_Lecture9

Block Size
حجم الكتلة

- عندما يتم استرداد كتلة من البيانات ووضعها في الذاكرة السريعة ، ليس فقط الكلمة المطلوبة يتم استردادها ولكن أيضا بعض الكلمات الملاصقة لها .
- كلما زاد حجم الكتلة إلى أحجام أكبر يتم جلب مزيد من البيانات المفيدة في الذاكرة السريعة ولكن احتمال استخدام بيانات مجلوبة حديثا يصبح أقل من احتمال إعادة استخدام بيانات التي لابد من استبدالها .

<u>هناك اعتباران يجب اخذهما فى الاعتبار :</u>

- كتل (Block) أكبر يخفض عدد الكتل (Block) التي تدرج في الذاكرة السريعة .
- كلما كبرت الكتلة (Block) أصبحت إي كلمة اضافية بعيدة عن الكلمة المطلوبة .

Dr. ramzi elghanuni_Lecture9

Number of Caches
عدد الذاكرة السريعة

- في البداية كان هناك مستوى واحد.
- مع تطور التقنية أصبح استخدام ذاكرة سريعة بمستويات متعددة.
- الذاكرة السريعة متعددة المستويات

- يمكن أن يكون المستوى الاول ملاصق للمعالج على الشريحة نفسها (ذاكرة سريعة داخلية).

- المستوى الثاني على اللوحة (ذاكرة سريعة خارجية).

- هناك استراتيجيتان في تصميم الذاكرة السريعة هل موحدة أم منفصلة.

- الموحدة تكون للبيانات والتعليمات معا.

- المنفصلة واحدة للتعليمات و اخرى للبيانات.

Dr. ramzi elghanuni_Lecture9